AFRL-IF-RS-TR-1998-71
Final Technical Report
May 1998

EXPERIMENTAL METHODS FOR EVALUATING PLANNING SYSTEMS

University of Massachusetts

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. A010

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-1998-71 has been reviewed and is approved for publication.

APPROVED:

WAYNE A. BOSCO
Project Engineer

FOR THE DIRECTOR:

NORTHRUP FOWLER, III, Technical Advisor
Information Technology Division
Information Directorate

EXPERIMENTAL METHODS FOR EVALUATING PLANNING SYSTEMS

Paul R. Cohen
Adele E. Howe

Contractor: University of Massachusetts
Contract Number: F30602-93-C-0100
Effective Date of Contract: July 1993
Contract Expiration Date: September 1997
Program Code Number: 61101E, 62301E, 63278F
Short Title of Work: Experimental Methods for Evaluating Planning Systems
Period of Work Covered: Jul 93 - Sep 97

Principal Investigator: Paul R. Cohen
Phone: (413) 545-3638
AFRL Project Engineer: Wayne Bosco
Phone: (315) 330-3578

Approved for public release; distribution unlimited.

This research was supported by the Defense Advanced Research Projects Agency of the Department of Defense and was monitored by Wayne A. Bosco, AFRL/IFTB, 525 Brooks Road, Rome, NY 13441-4505.

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.			
1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** May 1998	**3. REPORT TYPE AND DATES COVERED** Final Jul 93 - Sep 97	
4. TITLE AND SUBTITLE EXPERIMENTAL METHODS FOR EVALUATING PLANNING SYSTEMS			**5. FUNDING NUMBERS** C - F30602-93-C-0100 PE - 61101E, 62301E, 63278F PR - A010 TA - 00 WU - 01
6. AUTHOR(S) Paul R. Cohen, and Adele E. Howe			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Massachusetts Computer Science Department, Box 34610 Amherst MA 01003-4610			**8. PERFORMING ORGANIZATION REPORT NUMBER** N/A
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Defense Advanced Research Projects Agency 3701 North Fairfax Drive Arlington VA 22203-1714	Air Force Research Laboratory/IFTB 525 Brooks Road Rome NY 13441-4505		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER** AFRL-IF-RS-TR-1998-71
11. SUPPLEMENTARY NOTES Air Force Research Laboratory Project Engineer: Wayne A. Bosco/IFTB/(315) 330-3578			
12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution unlimited			**12b. DISTRIBUTION CODE**

13. ABSTRACT *(Maximum 200 words)*

The purpose of this research was to develop empirical tools that allow AI researchers to study their systems in a principled, systematic way. In the course of a three-year project the University of Massachusetts published over 30 refereed articles on empirical methods; a textbook, Empirical Methods for Artificial Intelligence; a web site devoted to research methods in AI and Computer Science; and over 40 papers at symposia, workshops, ad related forums.

This final report contains summaries of the first two years of the project and the final year. The project went according to plan with one significant deviation. The project didn't envision the power of the World Wide Web. Technical results were planned to be delivered in conventional ways by publishing software and reports. By the second year of the project, it became clear that a web site devoted to empirical methods was required. MIT Press granted permission to post excerpt of the recently published book. Technical articles were posted, along with more of the tutorial and resource material, onto the web site. The Evaluation of Intelligent Systems (EIS) web site is completed, and much of the technology developed under this program - including the CLASP system - can be found there. As the web site address is changing in the near future, we refer interested readers to the Experimental Knowledge Systems Laboratory web site (http://www-eksl.cs.umass.edu), from which pointers to EIS may be followed.

14. SUBJECT TERMS Empirical Tools, CLIP/CLASP, Multi-Stream Dependency Detection (MSDD), Multi-Event Dependency Detection (MEDD), Evaluation of Intelligent Systems (EIS), Assistant for Intelligent Data Exploration (AIDE)			**15. NUMBER OF PAGES** 44 **16. PRICE CODE**
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	**18. SECURITY CLASSIFICATION OF THIS PAGE** UNCLASSIFIED	**19. SECURITY CLASSIFICATION OF ABSTRACT** UNCLASSIFIED	**20. LIMITATION OF ABSTRACT** UL

Contents

Numerical Productivity Measures

- Refereed papers published: 35
- Refereed papers submitted: 1
- Invited papers published: 6
- Refereed workshop abstracts and symposia papers: 17
- Books or parts thereof published: 6
- Ph.D. dissertations: 3
- Unrefereed reports and articles: 14
- Invited presentations: 32
- Contributed presentations: 10
- Tutorials: 3
- Honors, including conference committees: 15
- Graduate students supported at least 25% time:7

Additional Researchers Working with the Principal Investigator

Subcontractor and Co-Investigator: Adele E. Howe

Graduate Students: Scott D. Anderson, Marc S. Atkin, Lisa A. Ballesteros, Dawn E. Gregory, J. Timothy Oates, Michael T. Rosenstein, Matthew D. Schmill, Robert St. Amant.

Other: David M. Hart (Lab Manager), David L. Westbrook (Systems Development Supervisor).

1 Executive Summary

The purpose of this research was to develop empirical tools that would allow AI researchers to study their systems in a principled, systematic way. The need was clear and well-documented [3]. In the course of this three-year project we have published over 30 refereed articles on empirical methods; a textbook, *Empirical Methods for Artificial Intelligence*; a web site devoted to research methods in AI and Computer Science; and over 40 papers at symposia, workshops, and related forums. Three graduate students have received PhD's in the course of this work. All are faculty members at good institutions (Adele Howe, Colorado State; Scott Anderson, Spelman College; Rob St. Amant, North Carolina State). And while we cannot establish a causal connection to our advocacy and publications, we are happy to report that many DARPA programs now have vigorous requirements and study groups on empirical evaluation. (For instance, Professor Cohen is principal consultant on evaluation in the HPKB program.) The tide is turning: methodology is recognized as important. Under the sponsorship of ARPA and Rome Laboratory, we have been leaders in the struggle.

This is our final report. It contains summaries of the first two years of the project and somewhat more detail on this, the final year. The project went according to plan, with one significant deviation. When we started the project we didn't envision the power of the World Wide Web (who did?). We planned to deliver our technical results in conventional ways, by publishing software and reports. By the second year of the project, however, it became clear that we should set up a web site devoted to empirical methods—especially as we had just published a book on the subject and MIT Press gave us permission to excerpt the book. So while we continued to publish technical articles and to improve the CLASP statistical package, we started to move more and more of the tutorial and resource material onto the web site. Now the Evaluation of Intelligent Systems (EIS) web site is nearly completed, and much of the technology developed under this program—including the CLASP system—can be found there. As the web site address is changing in the near future, we refer interested readers to the Experimental Knowledge Systems Laboratory web site (http://www-eksl.cs.umass.edu), from which pointers to EIS may be followed.

2 Summary of Technical Results and Accomplishments, Year One

We include a list of representative papers at the end of each section. These papers can be accessed at http://www-eksl.cs.umass.edu/publications.html, unless otherwise noted. Additional technical detail for Years One and Two are provided in our Annual Reports for those years.

2.1 Dependency Detection

We implemented the Dependency Detection (DD) module, which identifies significant dependencies among measured factors in execution traces of planner behavior. This module has been integrated with CLIP/CLASP for delivery to the CPE. We also developed a new algorithm for detecting dependencies in multiple concurrent streams, called Multi-Stream Dependency Detection (MSDD). This algorithm is particularly effective for analyzing time-series data such as multiple program execution traces [16].

Howe, A.E. 1993. Accepting the Inevitable: The Role of Failure Recovery in the Design of Planners. Ph.D. Thesis, University of Massachusetts/Amherst.

Howe, A.E. and Cohen, P.R. 1994. Detecting and Explaining Dependencies in Execution Traces. *Selecting Models from Data: Artificial Intelligence and Statistics IV*, P. Cheeseman and R.W. Oldford, Eds., pp. 161-182.

Howe, A.E. and Cohen, P.R. 1995. Understanding Planner Behavior. *Artificial Intelligence*, Special Issue on Planning Systems, Vol. 76, Nos. 1&2, pp. 125-166. Full text at http://satchmo.cs.colostate.edu:4936/upb/aij-final-tr.html.

Oates, T., Gregory, D.E. and Cohen, P.R. 1995. Detecting Complex Dependencies in Categorical Data. *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 417-423.

2.2 Path Analysis

We developed two causal induction algorithms for the Path Analysis (PA) module, which uses path analytic techniques to model observed planner behavior. Tests show that these algorithms outperform other leading causal induction algorithms [5].

Cohen, P.R., Gregory, D.E., Ballesteros, L.A. and St. Amant, R. 1995. Two Algorithms for Inducing Structural Equation Models from Data. *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 129-139.

Ballesteros, L.A. 1994. Regression-based Causal Induction with Latent Variable Mod-

els. *Proceedings of the Twelfth National Conference on Artificial Intelligence.* AAAI Press/The MIT Press, p. 1426.

2.3 Assistant for Intelligent Data Exploration (AIDE)

Development of the Assistant for Intelligent Data Exploration (AIDE) was begun to assist human analysts in exploratory data analysis, or EDA [2]. AIDE provides the Acausal Functional Relations (AFR) module functionality, assisting the knowledge engineer in discovering unanticipated but significant patterns in planner behavior. A prototype of AIDE was built and integrated with the Path Analysis and Dependency Detection modules.

St. Amant, R. and Cohen, P.R. 1995. Preliminary System Design for an EDA Assistant. *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 502-511.

2.4 Nonparametric Confidence Intervals (NCI)

We began initial work on the Nonparametric Confidence Intervals (NCI), or bootstrapping, module. Basic bootstrapping functionality is being implemented now, with more sophisticated forms to follow.

2.5 Case Studies Compilation

We successfully applied Dependency Detection and Path Analysis in tandem to analyze a significant aspect of planner behavior in PHOENIX [6]. This case study demonstrates the combined application of multiple experiment modules to explain a single but complex behavior pattern. Several other case studies we have developed illustrate the use of basic CLIP/CLASP tools and of individual modules, such as AFR.

Howe, A.E., St. Amant, R. and Cohen, P.R. 1994. Integrating Statistical Methods for Characterizing Causal Influences on Planner Behavior Over Time. *Proceedings of the Sixth IEEE International Workshop on Tools with Artificial Intelligence.* IEEE Computer Society Press, pp. 56-62.

2.6 Infrastructure Development

CLIP/CLASP Workshops. We conducted a very successful workshop on CLIP/CLASP at Rome Laboratory in October, 1993. The workshop was attended by Planning Initiative members from Pittsburgh, Rochester, Yale, GE, BBN, ISX, and Rome Laboratory.

Researchers were encouraged to bring their own simulation testbeds for hands-on application of CLIP/CLASP, and several did – Pittsburgh researchers brought TILEWORLD and Rochester researchers brought the TRAINS system.

In addition to the workshops, we continued to develop CLIP/CLASP, concentrating on three tasks: 1) maintaining and improving it for CPE users, 2) enhancing it to be a framework for the experiment modules, and 3) using it as a teaching vehicle to supplement Paul Cohen's textbook, *Empirical Methods for Artificial Intelligence.*

Anderson, S.D., Carlson, A., Westbrook D.S., Hart, D.M. and Cohen, P.R. 1993. Common Lisp Analytical Statistics Package: User Manual. Technical Report 93-55, Dept. of Computer Science, University of Massachusetts/Amherst.

Westbrook, D.L., Anderson, S.D., Hart, D.M. and Cohen, P.R. 1994. Common Lisp Instrumentation Package: User Manual. Technical Report 94-26, Dept. of Computer Science, University of Masschusetts/Amherst.

See also http://www-eksl.cs.umss.edu/research/clip-clasp-overview.html.

3 Summary of Technical Results and Accomplishments, Year Two

3.1 Further Development of Experiment Modules

Dependency Detection (DD) was extended to a multi-stream version (MSDD) and an incremental multi-stream version (IMSDD). MSDD was shown to be equal to many standard classification algorithms when run on UC Irvine Machine Learning datasets, even though it was not specifically designed as a classifier. However, a drawback of MSDD is that it works offline on batch data; IMSDD works online, using data as it becomes available. DD, MSDD and IMSDD are all integrated with CLASP through a common (but separate) interaction pane accessible from the CLASP main menu.

Schmill, M.D., Oates, T. and Cohen, P.R. 1995. Tools for Detecting Dependencies in AI Systems. *Proceedings of the 7th International IEEE Conference on Tools with Artificial Intelligence*, pp. 148-155.

Creating Agents that Use MSDD to Learn to Plan. We successfully applied MSDD to the task of automatically learning planning operators with context-dependent and probabilistic effects in environments where exogenous events change the state of the world. We were able to show two significant results:

- The number of search nodes required by MSDD to find these target operators scales approximately linearly with the size of the agent's state description, even though

the size of the operator space increases exponentially.

- The algorithm consistently returns small sets of operators that contain the target operators, as well as operators that capture structure that is implicit in the definition of the sample domain but that was not explicitly codified in the target operators.

Cohen, P.R. and Oates, T. 1995. Finding Structure in Streams. *Proceedings of the IDA-95 Symposium of the International Institute for Advanced Studies in Systems Research, Informatics and Cybernetics*, Vol. 1, pp. 27-31.

Oates, T. and Cohen, P.R. Searching for Planning Operators with Context-Dependent and Probabilistic Effects. *Proceedings of the 13th National Conference on Artificial Intelligence*, pp. 863-868.

Oates, T. and Cohen, P.R. Searching for Structure in Multiple Streams of Data. *Proceedings of the 13th International Conference on Machine Learning*, pp. 346-354.

Acausal Functional Relationship (AFR) or exploratory data analysis, module. We developed a new planning system that addresses the significant control issues inherent in exploratory data analysis. We also created a mixed-initiative user interface that allows the user to view the actions of the module and interact with the planner, asking for explanations and offering advice. This module developed into a full-fledged system with its own interface.

St. Amant, R. and Cohen, P.R. 1 1996. A Planner for Exploratory Data Analysis. *Proceedings of the 3rd International Conference on Artificial Intelligence Planning Systems*, pp. 205-212.

St. Amant, R. and Cohen, P.R. 1995. A Case Study of Planning for Exploratory Data Anlysis. *Proceedings of the IDA-95 Symposium of the International Institute for Advanced Studies in Systems Research, Informatics and Cybernetics*, Vol. 1, pp. 1-5.

Nonparametric Confidence Intervals (NCI), or bootstrapping, module. Bootstrapping functionality was added to the CLASP interaction pane, providing the menu-driven capability to bootstrap most of the common CLASP tests and statistics.

Piater, J.H. and Cohen, P.R. 1997. A Randomized ANOVA Procedure for Comparing Performance Curves. Technical Report 97-41, Dept. of Computer Science, University of Massachusetts/Amherst.

Path Analysis (PA), or causal modeling, module. The causal modeling algorithms developed in year one were integrated into CLASP through a special purpose, PA Interface accessible from the CLASP main menu.

Cohen, P.R., Gregory, D.E. Ballesteros, L.A. and St. Amant, R. 1995. Two Algorithms for Inducing Structural Equation Models from Dta. *Preliminary Papers of the 5th International Workshop on Artificial Intelligence and Statistics*, pp. 129-139.

3.2 Case Studies Compilation

- Publication of textbook, *Empirical Methods for Artificial Intelligence*, incorporating many of the methods and case studies being developed in this project.
- Use of empirical methods to identify serious search control difficulties with a well-known, contemporary AI planner, UCPOP, and to investigate possible solutions. This work by the subcontractor (Howe) provides a highly visible demonstration of the empirical approach.
- Also at CSU, empirical methods were used to begin development of an information gathering agent for querying WWW search engines. This agent uses techniques from machine learning and planning to learn how to formulate search plans based on past query results. Each plan employs a subset of available web search engines that have been shown to be effective based on similar queries in the past.

Cohen, P.R. 1995. *Empirical Methods for Artificial Intelligence.* Cambridge: The MIT Press.

Dreilinger, D. and Howe, A.E. 1996. An Information Gathering Agent for Querying Web Search Engines. Colorado State University, Computer Science Department Technical Report 96-111.

Srinivasan, R. and Howe, A.E. 1995. Comparison of Methods for Improving Search Efficiency in a Partial-Order Planner. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1620-1626.

3.3 Infrastructure Development

- Integration of CLIPCLASP and four Experiment Modules into a comprehensive empirical analysis environment. This environment eventually included additional tools for building real-time simulators and for mixed-initiative planning of experiments.
- CLASP for the Macintosh. A version of CLASP with a native Macintosh interface was developed for distribution with the *Empirical Methods* textbook, and many of the text's exercises use CLASP for the Macintosh.

See http://www-eksl.cs.umass.edu/research/clip-clasp-updates.html.

4 Summary of Technical Results and Accomplishments, Year Three

The third and final year of the project has seen significant progress on five fronts:

- CLASPWEB. The CLASP system is a statistics package written in Common Lisp, easily intergrated into AI development environments through the CLIP instrumentation package. As a standalone system, CLASP provides the functionality of most data analysis packages. CLASPWEB is, as far as we can tell, the first data analysis package that can be run over the Web. CLASP resides on a server, clients submit data, and analyses are returned. See `http://satchmo.cs.colostate.edu:4936` for access, although the server will change in the near future.
- Dependency Detection (DD) Implementation. The Dependency Detection algorithm designed by Professors Howe and Cohen has been improved in several ways. See http://satchmo.cs.colostate.edu:
4936/upb/node3.html.
- Multi-Stream Dependency Detection (MSDD). Whereas Dependency Detection finds structure in a single data stream, MSDD finds structure in multiple streams. The algorithm has been improved and parallelized.
- Overfitting. One of the most pernicious and ubiquitous pathologies of induction algorithms and some search algorithms is overfitting. Basically, overfitting means that an algorithm goes too far: It may add too much structure to a model, or it may learn things that are more complex than they should be, or it may extend search farther than it should. We have provided a statistical theory of overfitting and simple statistical methods to avoid it.
- Evaluation of Intelligent Systems Website. The EIS Web site was set up to provide one-stop shopping for empirically-inclined researchers. It is the delivery vehicle for much of the technology developed under this program.

We will discuss these contributions in more detail.

4.1 CLASPWeb

To our knowledge, CLASPWEB is the first data analysis package to run on the World Wide Web. It is a web-based implementation of the Common Lisp Analytical Statistics Package, developed under this contract. CLASP itself is widely used, for example it is available with Prof. Cohen's book *Empirical Methods for Artificial Intelligence*. CLASP runs on the Macintosh and Unix boxes, but the latter implementation is difficult in some

respects and requires a Lisp license. It became clear to us that browsers like Netscape provide platform independent interfaces to software packages, hence CLASPWEB.

The major technical problems in CLASPWEB have to do with the memoryless nature of web protocols. When a client requests a page from a server, it is sent and the connection is terminated. No information about the client persists on the server (although cookies are changing that to some extent). So imagine the plight of a data analyst who requests a statistical operation from the CLASPWEB server: He must ship his entire dataset to the server every time. This creates web traffic and takes time. The solution we adopted was to allow a user to store his or her dataset on the CLASPWEB server for a period of time. Then the client merely requests operations on the dataset from the server, and the results are shipped back to the client.

So far, so good, but some operations have side effects. For example, partitioning a dataset creates several other datasets. Transforming a variable creates another variable. While some statistical operations (e.g., taking the mean) are *reducing* in the sense that they reduce data to something smaller — a statistic — many operations create new data. Where is it stored? Currently, on the CLASPWEB server. This is not a realistic long term solution.

A related problem is security. Some CLASP operations allow the user to write and execute arbitrary Lisp code. Clearly, these operations cannot be allowed on CLASPWEB, otherwise clients would be able to make the server do arbitrary things.

The current status of CLASPWEB is that it is on the Web, used sparingly, and continually under development. We expect it to be more heavily used as the full CLASP functionality becomes available.

4.2 Dependency Detection

Dependency detection refers to algorithms that find dependencies betwen discrete events in time series. For instance, when one event follows another more often that would be expected by chance, we say they are linked by a dependency. Professors Howe and Cohen invented dependency detection before this contract began and they have been refining it since. The most recent development is from Professor Howe's laboratory. Dependency detection finds little "nuggets" in time series – pairs of events, or triples, but not longer chains of events. In particular, dependency detection doesn't allow one to infer a finite state machine from the dependencies it finds. Professor Howe's most recent work fixes this problem. She has an algorithm that merges dependencies between pairs of events to get a finite state machine that generates a time series of events. This work is reported in [7].

4.3 Multi Stream Dependency Detection

Finding patterns in event-based data is an important and difficult problem that has received little attention from researchers in either AI or statistics. We have developed and tested a family of algorithms for finding structure in multivariate time series [11,13,14].

One application of these techniques is to predicting events in computer networks. Networks produce large amounts of event-based data, and management of such networks is largely driven by the generation and interpretation of events. A problem that plagues network managers is the large number of events of different types from disparate locations in the network that result from network faults [10]. Finding patterns in those events to form clusters of related events is important for reducing the amount of information that must be interpreted and for understanding the state of the network (e.g. by identifying whether a set of events represents the effects of a single fault or multiple, concurrent faults). A version of MSDD called called Multi-Event Dependency Detection (MEDD). MEDD finds *dependencies* between patterns of events recorded in event logs.

Another recent development is the parallelization of MSDD. Experiments that are currently underway suggest that the search space of MSDD can be efficiently divided among several machines, reducing search times. We do not yet know by what factor runtimes are reduced. Clearly, communication overhead pervents us getting a reduction of runtime to $1/p$ on p processors, but the deviation from this ideal appears to be small, on the basis of preliminary results.

We are exploring commercial possibilities for MSDD.

4.4 Overfitting

Overfitting is a widely observed pathology of induction algorithms. Overfitted models contain unnecessary structure that reflects nothing more than random variation in the data sample used to construct the model. Portions of these models are literally wrong, and can mislead users. Overfitted models are less efficient to store and use than their correctly-sized counterparts. Finally, overfitting can reduce the accuracy of induced models on new data [8,15].

We have discovered why induction algorithms overfit. The reason is quite subtle: Algorithms generally don't account for the fact that the components they are considering adding to models are the *best* of several components. We proved that the maximum component has statistical distributions quite unlike ordinary components, resulting in systematic overestimates of the accuracy of components. We have shown empirically the importance of accounting for multiple comparisons when evaluating models (Jensen and Cohen, 1997). We have also invented several techniques to control overfitting. Based on experiments with artificial and realistic datasets, *Bonferroni pruning* produces trees that are smaller and at least as accurate as trees pruned using several other common

approaches.

We also have shown that the techniques used by other researchers to compensate for overfitting don't really work. Often such techniques involve *data reduction*, the removal of training instances prior to tree construction. For example, some techniques identify instances that are "bad" and remove them from the training set, while others actively build a training set from available instances by selecting those that are "good". Whether the explicit goal of any given technique is increased accuracy or smaller trees, the latter is invariably observed. John's ROBUST-C4.5 treats misclassified training instances as outliers, iteratively removing them and building a new tree [9]. The result over a large number of datasets is trees that are much smaller than those built by C4.5, but that have roughly equivalent accuracy. We demonstrated that *any* technique that throws away data will have this effect: There's nothing particularly smart about the techniques in the literature. Random discarding of data would accomplish the same thing [12].

4.5 Evaluation of Intelligent Systems

Unlike Biology, Medicine, Psychology, etc., Computer Science has no curriculum in Research Methods. We need one. Our survey of 150 papers at AAAI-90 showed lots of theoretical AI and system-building, but only 10% merged theoretical and empirical work in experiments, the hallmark of science. In a recent issue of the AI Journal on Empirical AI, which we edited, half the accepted papers (and all the rejected ones) needed methodological revision. Some basic errors are endemic and very damaging. For example, using the wrong distributions in decision tree induction causes overfitting, multiple pairwise t tests underestimate error probabilities, huge samples obscure causes of variance, and so on. The good news is that these problems are easy to understand and fix, so improvements in practice could be global and fast given a global, fast way to get the word out.

The Evaluation of Intelligent Systems (EIS) web site, sponsored by Rome Laboratory, provides information on exploratory data analysis, visualization, and discovery procedures; on experiment design, control conditions and common pitfalls; on statistical hypothesis testing, including new Monte Carlo methods; on parameter estimation and modeling; on common experimental methods for evaluating systems and for probing them to find causes of behaviors.

EIS also provides software for data analysis (CLASPWEB, see above). It provides links to datasets, analytical algorithms, newsgroups, and reports of new techniques. It is anchored by *Empirical Methods for Artificial Intelligence* [4]. EIS is an online, interactive curriculum, handbook, field-guide, analysis package and reference library for empirical methods.

Most notably, EIS has become the delivery vehicle for many of the algorithms, techniques, and publications that we have developed under the this contract.

Howe, Adele E. and G. Somlo. 1997. Modeling Discrete Event Sequences as State Transition Diagrams. In *Advances in Intelligent Data Analysis: Reasoning About Data.* Proceedings of IDA-97, pp. 573-584.

Howe, Adele E. 1996. Detecting Imperfect Patterns in Event Streams Using Local Search. In *Learning from Data: Artificial Intelligence and Statistics V*, D. Fisher and H. Lenz (Eds.). Springer-Verlag, pp. 155-164.

Oates, Tim, Matthew Schmill, David Jensen and Paul R. Cohen. 1997. A Family of Algorithms for Finding Temporal Structure in Data. In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 371-378.

Oates, Tim, Matthew D. Schmill, Dawn E. Gregory and Paul R. Cohen. 1996. Detecting Complex Dependencies in Categorical Data. In *Learning from Data: Artificial Intelligence and Statistics V*, D. Fisher and H. Lenz (Eds.). Springer-Verlag, pp. 185-195.

Jensen, D. and Cohen, P.R. Multiple Comparisons in Induction Algorithms. Submitted to *Machine Learning.*

Jensen, D., Oates, T. and Cohen, P.R. 1997. Building Simple Models: A Case Study with Decision Trees. In *Advances in Intelligent Data Analysis: Reasoning About Data.* Proceedings of IDA-97, pp. 211-222.

Cohen, Paul R. and David Jensen. 1997. Overfitting Explained. In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 115-122.

5 Publications, Reports and Articles

5.1 Books, or Parts Thereof

Howe, Adele E. 1996. Detecting Imperfect Patterns in Event Streams Using Local Search. In *Learning from Data: Artificial Intelligence and Statistics V*, D. Fisher and H. Lenz (Eds.). Springer-Verlag, pp. 155-164.

Oates, Tim, Matthew D. Schmill, Dawn E. Gregory and Paul R. Cohen. 1996. Detecting Complex Dependencies in Categorical Data. In *Learning from Data: Artificial Intelligence and Statistics V*, D. Fisher and H. Lenz (Eds.). Springer-Verlag, pp. 185-195.

St. Amant, Robert and Paul R. Cohen. 1996. Control Representation in an EDA Assistant. In *Learning from Data: Artificial Intelligence and Statistics V*, D. Fisher and H. Lenz (Eds.). Springer-Verlag, pp. 353-362.

Cohen, Paul R. 1995. *Empirical Methods for Artificial Intelligence.* Cambridge: The MIT Press.

Cohen, Paul R., David M. Hart, Robert St. Amant, Lisa A. Ballesteros and Adam Carlson. 1994. Path Analysis Models of an Autonomous Agent in a Complex Environment. In *Selecting Models from Data: AI and Statistics IV*, P. Cheeseman and R. W. Oldford (Eds.). Springer Verlag, pp. 243-251.

Howe, Adele E. and Paul R. Cohen. 1994. Detecting and Explaining Dependencies in Execution Traces. In *Selecting Models from Data: AI and Statistics IV*, P. Cheeseman and R. W. Oldford (Eds.) Springer-Verlag, pp. 71-77.

5.2 Journal Articles

Dreilinger, D. and Adele E. Howe. Experiences with Selecting Search Engines Using Meta-Search. To appear in *ACM Transactions on Information Systems.*

Anderson, Scott D. and Paul R. Cohen. 1996. Timed Common LISP: The Duration of Deliberation. In *SIGART Bulletin*, Vol. 7, No. 2, pp. 11-15.

Atkin, Marc and Paul R. Cohen. 1996. Monitoring Strategies for Embedded Agents: Experiments and Analysis. In *Journal of Adaptive Behavior.* Vol. 4, No. 2, pp. 125-172.

Cohen, Paul R. 1996. Getting What You Deserve from Data. In *IEEE Expert, Intelligent Systems and Their Applications*, Vol. 11, No. 5, pp. 12-14.

Anderson, Scott D., David M. Hart, David L. Westbrook and Paul R. Cohen. 1995. A Toolbox for Analyzing Programs. In *International Journal on Artificial Intelligence Tools.* Vol. 4, Nos. 1&2, pp. 257-279.

Howe, Adele E. and Paul R. Cohen. 1995. Understanding Planner Behavior. In *Artificial Intelligence*, Special Issue on Planning Systems. Vol. 76, Nos. 1&2, pp. 125-166.

Howe, Adele E. 1995. Improving the Reliability of AI Planning Systems by Analyzing Their Failure Recovery. In *IEEE Transactions on Knowledge and Data Engineering*, Special Issue on Dependability of AI Systems.

Hanks, Steve, Martha E. Pollack and Paul R. Cohen. 1993. Benchmarks, Testbeds, Controlled Experimentation, and the Design of Agent Architectures. In *AI Magazine*, Vol. 14, No. 4, pp. 17-42.

5.3 Refereed Conference Papers

Gregory, Dawn E. and Paul R. Cohen. 1997. Integrating Many Techniques for Discovering Structure in Data. In *Advances in Intelligent Data Analysis: Reasoning About Data.* Proceedings of IDA-97, pp. 231-238.

Howe, Adele E. and G. Somlo. 1997. Modeling Discrete Event Sequences as State Transition Diagrams. In *Advances in Intelligent Data Analysis: Reasoning About Data.* Proceedings of IDA-97, pp. 573-584.

Jensen, David, Tim Oates and Paul R. Cohen. 1997. Building Simple Models: A Case Study with Decision Trees. In *Advances in Intelligent Data Analysis: Reasoning About Data.* Proceedings of IDA-97, pp. 211-222.

McGeoch, C.C., Doina Precup and Paul R. Cohen. 1997. How to Find Big-Oh in Your Data Set (and How Not to). In *Advances in Intelligent Data Analysis: Reasoning About Data.* Proceedings of IDA-97, pp. 41-52.

St. Amant, Robert and Paul R. Cohen. 1997. Evaluation of a Semi-Autonomous Assistant for Exploratory Data Analysis. In *Proceedings of the First International Conference on Autonomous Agents*, pp. 355-362.

St. Amant, Robert and Paul Cohen. Interaction with a Mixed Initiative System for Exploratory Data Analysis. In *Proceedings of the Third International Conference on Intelligent User Interfaces*, pp. 15-22.

Howe, Adele E. and L.D. Pyeatt. 1996. Constructing Transition Models of AI Planner Behavior. In *Proceedings of the 11th Knowledge-Based Software Engineering Conference.*

Oates, Tim and Paul R. Cohen. 1996. Searching for Planning Operators with Context-Dependent and Probabilistic Effects. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence*, pp. 863-868.

Oates, Tim and Paul R. Cohen. 1996. Searching for Structure in Multiple Streams of Data. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 346-354.

St. Amant, Robert and Paul R. Cohen. 1996. A Planner for Exploratory Data Analysis. In *Proceedings of the Third International Conference on Artificial Intelligence Planning Systems*, pp. 205-212.

Cohen, Paul R. and Tim Oates. Finding Structure in Streams. In *Advances in Intelligent Data Analysis: Proceedings of IDA-95*, pp. 27-31.

Gregory, Dawn E. and Paul R. Cohen. 1995. A Function Modeling Approach to Empirical Science. In *Book of Abstracts of the 10th International Conference on Mathematical and Computer Modeling and Scientific Computing*, p. 42.

Schmill, Matthew D., Tim Oates and Paul R. Cohen. 1995. Tools for Detecting Dependencies in AI Systems. In *Proceedings of the Seventh International IEEE Conference on Tools with Artificial Intelligence*, pp. 148-155.

Srinivasan, R. and Adele E. Howe. 1995. Comparison of Methods for Improving Search Efficiency in a Partial-Order Planner. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1620-1626.

St. Amant, Robert, Yoshitaka Kuwata and Paul R. Cohen. 1995. Monitoring Progress with Dynamic Programming Envelopes. In *Proceedings of the Seventh International IEEE Conference on Tools with Artificial Intelligence*, pp. 426-433.

Anderson, Scott D., Adam Carlson, David L. Westbrook, David M. Hart and Paul R. Cohen. 1994. Tools for Experiments in Planning. In *Proceedings of the Sixth IEEE International Conference on Tools with AI.* IEEE Computer Society Press, pp. 615-623.

Atkin, Marc S. and Paul R. Cohen. 1994. Learning Monitoring Strategies: A Difficult Genetic Programming Application. In *Proceedings, IEEE International Conference on Evolutionary Computation.* Piscataway NJ: IEEE, pp. 328-332.

Cohen, Paul R., Marc S. Atkin and Eric A. Hansen. 1994. The Interval Reduction Strategy for Monitoring Cupcake Problems. In *Proceedings, From Animals to Animats, the Third International Conference on Simulation of Adaptive Behavior.* MIT Press, pp. 82-90.

Hansen, Eric A. 1994. Cost-Effective Sensing During Plan Execution. In *Proceedings of the Twelfth National Conference on Artificial Intelligence.* AAAI Press/MIT Press, pp. 1029-1035.

Howe, Adele E. and Aaron Fuegi. 1994. How Would We Know if a Program Failure was Caused by What It Did Yesterday? In *Proceedings, Sixth IEEE International Conference on Tools with Artificial Intelligence.*

Howe, Adele E., Robert St. Amant and Paul R. Cohen. 1994. Integrating Statistical Methods for Characterizing Causal Influences on Planner Behavior Over Time. In *Proceedings, Sixth IEEE International Conference on Tools with Artificial Intelligence.*

Oates, Tim and Paul R. Cohen. 1994. Toward a Plan Steering Agent: Experiments with Schedule Maintenance. *Proceedings, Second International Conference on Artificial Intelligence Planning Systems.* AAAI Press, pp. 134-139.

St. Amant, Robert and Paul R. Cohen. 1994. Automated Analysis of Complex Data. In *Proceedings of the Ninth Annual Goddard Conference on Space Applications of Artificial Intelligence*, pp. 161-172.

Cohen, Paul R., Adam Carlson, and Lisa Ballesteros. 1993. Automating Path Analysis for Building Causal Models from Data. In *Proceedings of the Tenth International Conference on Machine Learning.* Morgan Kaufmann Publishers, Inc., pp. 57-64.

Cohen, Paul R., Robert St. Amant, and David M. Hart. 1992. Early Warnings of Plan Failure, False Positives and Envelopes: Experiments and a Model. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.* Lawrence Earlbaum Associates, Inc., pp. 773-778.

Hart, David M. and Paul R. Cohen. 1992. Predicting and Explaining Success and Task Duration in the Phoenix Planner. In *Artificial Intelligence Planning Systems: Proceedings of the First International Conference (AIPS92).* Morgan Kaufmann Publishers, Inc., pp. 106-115.

Howe, Adele E. 1992. Analyzing Failure Recovery to Improve Planner Design. In *Proceedings of the Tenth National Conference on Artificial Intelligence.* AAAI Press/MIT Press, pp. 387-392.

5.4 Invited Papers

Cohen, Paul R., Tim Oates and Robert St. Amant. 1996. Plan Steering and Mixed-Initiative Planning. In *Advanced Planning Technology: Technological Achievements of the ARPA/Rome Laboratory Planning Initiative*, A. Tate, Ed. AAAI Press, pp. 105-112.

Anderson, Scott D., Adam Carlson, David L. Westbrook, David M. Hart and Paul R. Cohen. 1994. Tools for Experiments in Planning. In *Proceedings, ARPA/Rome Laboratory Planning Initiative Workshop*, M.H. Burstein, Ed. Morgan Kaufmann Publishers, Inc., pp. 423-432.

Cohen, Paul R. and Marc S. Atkin. 1994. The Interval Reduction Strategy for Monitoring Cupcake Problems. In *Proceedings, ARPA/Rome Lab Planning Initiative Workshop*, M.H. Burstein, Ed. Morgan Kaufmann Publishers, Inc., pp. 15-26.

Cohen, P. R., T. Dean, Y. Gil, M. Ginsberg and L. Hoebel. 1994. Handbook of Evaluation for the ARPA/Rome Lab Planning Initiative. In *Proceedings, ARPA/Rome Laboratory Planning Initiative Workshop*, M.H. Burstein, Ed. Morgan Kaufmann Publishers, Inc., pp. 519-536.

Howe, Adele E., Robert St. Amant and Paul R. Cohen. 1994. Statistical Methods for Characterizing Causal Influences on Planner Behavior Over Time. In *Proceedings, ARPA/Rome Laboratory Planning Initiative Workshop*, M.H. Burstein, Ed. Morgan Kaufmann Publishers, Inc., pp. 445-452.

Oates, Tim and Paul R. Cohen. 1994. Mixed-Initiative Schedule Maintenance: A First Step Toward Plan Steering. In *Proceedings, ARPA/Rome Lab Planning Initiative Workshop*, M.H. Burstein, Ed. Morgan Kaufmann Publishers, Inc., pp. 133-143.

5.5 Refereed Workshop Abstracts and Symposia Papers

Cohen, Paul R. and David Jensen. 1997. Overfitting Explained. In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 115-122.

Gregory, Dawn E. and Paul R. Cohen. 1997. Intelligent Assistance for Computational Scientists: Integrated Modeling, Experimentation, and Analysis. In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 231-238.

McGeoch, C.C. and Paul R. Cohen. 1997. How to Find Big-Oh in Your Data Set (and How Not To). In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 347-354.

Oates, Tim and David Jensen. 1997. The Effects of Training Set Size on Decision Tree Complexity. In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 379-390.

Oates, Tim, Matthew Schmill, David Jensen and Paul R. Cohen. 1997. A Family of Algorithms for Finding Temporal Structure in Data. In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 371-378.

St. Amant, Robert and Paul R. Cohen. 1997. Building an EDA Assistant: A Progress Report. In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 501-512.

Gregory, Dawn E., Lixin Gao, Arnold L. Rosenberg and Paul R. Cohen. 1996. An Empirical Study of Dynamic Scheduling on Rings of Processors. In *Proceedings of the Eighth IEEE Symposium on Parallel and Distributed Processing*, pp. 470-473.

Oates, Tim and Paul R. Cohen. 1996. Learning Planning Operators with Conditional and Probabilistic Effects. In *Working Notes of the AAAI-96 Spring Symposium on Planning with Incomplete Information for Robot Problems*, pp. 86-94.

Anderson, Scott D. and Paul R. Cohen. 1995. Segregating Planners and Their Environments. In *Working Notes of the Symposium on Integrated Planning Applications*, AAAI-95 Spring Symposium Series.

Anderson, Scott D., Adam Carlson, David L. Westbrook, David M. Hart and Paul R.

Cohen. 1995. Tools for Empirically Analyzing AI Programs. In *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 35-41.

Cohen, Paul R., Dawn E. Gregory, Lisa A. Ballesteros and Robert St. Amant. 1995. Two Algorithms for Inducing Structural Equation Models from Data. In *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 129-139.

Howe, Adele E. 1995. Finding Dependencies in Event Streams Using Local Search. In *Preliminary Papers of the Fifth International Workshop on AI and Statistics*, pp. 271-277.

Oates, Tim, Dawn E. Gregory and Paul R. Cohen. 1995. Detecting Complex Dependencies in Categorical Data. In *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 417-423.

Atkin, Marc S. and Paul R. Cohen. 1993. Genetic Programming to Learn an Agent's Monitoring Strategy. In *Working Notes of the Learning Action Models Workshop*, Eleventh National Conference on Artificial Intelligence, Washington DC, pp. 36-41.

Cohen, Paul R. and David M. Hart. 1993. Path Analysis Models of an Autonomous Agent in a Complex Environment. In *Preliminary Papers of the Fourth International Workshop on AI and Statistics*, pp. 185-189.

Cohen, Paul R., Lisa Ballesteros, Adam Carlson, and Robert St. Amant. 1993. Automating Path Analysis for Building Causal Models from Data: First Results and Open Problems. In *Working Notes of the Knowledge Discovery in Databases Workshop*, Eleventh National Conference on Artificial Intelligence, Washington DC, pp. 153-161.

Hansen, Eric A. and Paul R. Cohen. 1993. Learning Monitoring Strategies to Compensate for Model Uncertainty. In *Working Notes of the Learning Action Models Workshop*, Eleventh National Conference on Artificial Intelligence, Washington DC, pp. 33-35.

5.6 Ph.D. Dissertations

St. Amant, Robert. August 1996. A Mixed-Initiative Planning Approach to Exploratory Data Analysis.

Anderson, Scott D. August 1995. A Simulation Substrate for Real-Time Planning.

Howe, Adele E. February 1993. Accepting the Inevitable: The Role of Failure Recovery in the Design of Planners.

5.7 Unrefereed Reports and Articles

Oates, Tim, David Jensen and Paul R. Cohen. 1997. Automatically Acquiring Rules for Event Correlation from Event Logs. Technical Report 97-14, Dept. of Computer Science, University of Masschusetts/Amherst.

Piater, Justus and Paul R. Cohen. 1997. A Randomized ANOVA Procedure for Comparing Performance Curves. Technical Report 97-41. Dept. of Computer Science, University of Massachusetts/Amherst.

Dreilinger, D. and Adele E. Howe. 1996. An Information Gathering Agent for Querying Web Search Engines. Colorado State University, Computer Science Technical Report 96-111.

Oates, Tim, Matthew D. Schmill and Paul R. Cohen. 1996. Parallel and Distributed Search for Structure in Multivariate Time Series. Technical Report 96-23, Dept. of Computer Science, University of Massachusetts/Amherst.

Atkin, Marc S. and Paul R. Cohen. 1995. Monitoring in Embedded Agents. Technical Report 95-66, Dept. of Computer Science, University of Massachusetts/Amherst.

Oates, Tim and Paul R. Cohen. 1995. Evaluation of a Mixed-Initiative Approach to Schedule Maintenance. Technical Report 95-08, Dept. of Computer Science, University of Massachusetts/Amherst.

Schmill, Matthew D. and Paul R. Cohen. 1995. Learning Predictive Generalizations for Multiple Streams: An Incremental Algorithm. Technical Report 95-55, Dept. of Computer Science, University of Massachusetts/Amherst.

Srinivasan, R. and Adele E. Howe. 1995. New Methods for Plan Selection and Refinement in a Partial Order Planner. Colorado State University, Computer Science Technical Report CS-95-03.

Cohen, Paul R., Dawn E. Gregory and Robert St. Amant. 1994. Regression Can Build Predictive Causal Models. Technical Report 94-15, Dept. of Computer Science, University of Massachusetts/Amherst.

Gregory, Dawn E. and Paul R. Cohen. 1994. Toward a Scientist's Empirical Assistant. Technical Report 94-85, Dept. of Computer Science, University of Massachusetts/Amherst.

Oates, Tim and Paul R. Cohen. 1994. Humans Plus Agents Maintain Schedules Better than Either Alone. Technical Report 94-03. Department of Computer Science, University of Massachusetts/Amherst.

Oates, Tim. 1994. MSDD as a Tool for Classification. EKSL Memo #94-29, Department of Computer Science, University of Massachusetts/Amherst.

Westbrook, David L., Scott D. Anderson, David M. Hart and Paul R. Cohen. 1994. Com-

mon Lisp Instrumentation Package: User Manual. Technical Report 94-26. Department of Computer Science, University of Massachusetts/Amherst.

Anderson, Scott D., Adam Carlson, David L. Westbrook, David M. Hart, and Paul R. Cohen. 1993. Common Lisp Analytical Statistics Package: User Manual. Technical Report 93-55. Department of Computer Science, University of Massachusetts/Amherst.

5.8 Conferences, Workshops and Presentations

5.8.1 Invited Presentations

August 1997: "Integrating Many Techniques for Discovering Structure in Data," presented by Paul Cohen at IDA-97, London UK.

August 1997: "Modeling Discrete Event Sequences as State Transition," presented by Adele Howe at IDA-97, London UK.

August 1997: "Building Simple Models: A Case Study with Decision Trees," presented by Paul Cohen at IDA-97, London UK.

August 1997: "How to Find Big-Oh in Your Data Set (and How Not to)," presented by Paul Cohen at IDA-97, London UK.

February 1997: "Evaluation of a Semi-Autonomous Assistant for Exploratory Data Analysis," presented by Robert St. Amant at the First International Conference on Autonomous Agents, Marina del Rey CA.

January 1997: "Building an EDA Assistant: A Progress Report," presented by Robert St. Amant at the Sixth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale FL.

January 1997: "The Effects of Training Set Size on Decision Tree Complexity," presented by Tim Oates at the Sixth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale FL.

January 1997: "A Family of Algorithms for Finding Temporal Structure in Data," presented by Matthew Schmill at the Sixth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale FL.

January 1997: "How to Find Big-Oh in Your Data Set (and How Not To)," presented by Paul Cohen at the Sixth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale FL.

January 1997: "Intelligent Assistance for Computational Scientists: Integrated Modeling, Experimentation, and Analysis," presented by Dawn Gregory at the Sixth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale FL.

January 1997: "Overfitting Explained," presented by David Jensen at the Sixth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale FL.

January 1997: "Interaction with a Mixed Initiative System for Exploratory Data Analysis," presented by Robert St. Amant at the Third International Conference on Intelligent User Interfaces, Orlando FL.

September 1996: "Constructing Transition Models of AI Planner Behavior," presented by Adele Howe at the Eleventh Knowledge-Based Software Engineering Conference, Syracuse NY.

August 1996: "Searching for Planning Operators with Context-Dependent and Probabilistic Effects," presented by Tim Oates at the Thirteenth International Conference on Artificial Intelligence, Stanford University.

July 1996: "Searching for Structure in Multiple Streams of Data," presented by Tim Oates at the Thirteenth International Conference on Machine Learning, Bari, Italy.

May 1996: "A Planner for Exploratory Data Analysis," presented by Robert St. Amant at the Third International Conference on Artificial Intelligence Planning Systems, Edinburgh, Scotland.

May 1996: "Plan Steering and Mixed-Initiative Planning," presented by Paul Cohen at ARPI Workshop, Edinburgh, Scotland.

May 1996: "Experimental Methods for Evaluating Planning Systems: Progress Report," presented by Adele Howe at ARPI Workshop, Edinburgh, Scotland.

April 1996: "Evaluation as a Basis for AI Software Development," presented by Adele Howe at the University of Chicago, Chicago IL.

November 1995: "Tools for Detecting Dependencies in AI Systems," presented by Matthew Schmill at the Seventh International IEEE Conference on Tools with Artificial Intelligence, Washington DC.

November 1995: "Monitoring Progress with Dynamic Programming Envelopes," presented by Robert St. Amant at the Seventh International IEEE Conference on Tools with Artificial Intelligence, Washington DC.

August 1995: "The Duration of Deliberation," presented by Scott Anderson at the IJCAI-95 Workshop on Anytime Algorithms and Deliberation Scheduling, Montreal, Canada.

August 1995: "A Case Study of Planning for Exploratory Data Analysis," presented by Paul Cohen at IDA-95, Baden-Baden, Germany.

August 1995: "Finding Structure in Streams," presented by Paul Cohen at IDA-95, Baden-Baden, Germany.

July 1995: "A Function Modelling Approach to Empirical Science," presented by Dawn

Gregory at the Tenth International Conference on Mathematical and Computer Modelling and Scientific Computing, Boston MA.

March 1995: "Segregating Planners and Their Environments," presented by Scott Anderson at the AAAI-95 Spring Symposium on Integrated Planning Applications, Stanford University.

March 1995: "Preliminary System Design for an EDA Assistant," presented by Robert St. Amant at the AAAI-95 Spring Symposium on Systematic Methods of Scientific Discovery, Stanford University.

October 1994: "When Good Planners Go Bad," presented by Adele Howe to the Math and Computer Science Dept., Colorado School of Mines, Golden CO.

June 1994: "Using Statistics in AI," presented by Paul Cohen at the *Ninth Annual Conference: Making Statistics More Effective in Schools of Business*, Rutgers University Graduate School of Management.

March 1994: "Characterizing Causal Influences on Planner Behavior to Assist Debugging," presented by Adele Howe at SRI, Menlo Park CA.

February 1994: "Evaluation in the Planning Initiative," presented by Paul Cohen at the *ARPA/Rome Laboratory Knowledge-Based Planning and Scheduling Initiative Workshop*, Tucson, AZ.

5.9 Contributed Presentations

September 1995: Paul Cohen presented a talk at Rome Laboratory on Intelligent Data Analysis.

November 1994: Paul Cohen presented a talk at Rome Laboratory on Dependency Detection.

February 1994: Paul Cohen, David Hart, David Westbrook (UMass) and Adele Howe (Colorado St. Univ.) attended the ARPA/Rome Laboratory Knowledge-Based Planning and Scheduling Initiative Workshop in Tucson, AZ and participated in the following:

- A panel led by Paul Cohen describing the "Handbook of Evaluation for the ARPA/Rome Lab Planning Initiative."
- A demo led by Hart and Westbrook of our interactive plan steering system (Oates et al. 1994).
- A "CLIP/CLASP Workshop" led by Cohen and Westbrook.

December 1993: Paul Cohen attended the Quarterly Planning Initiative meeting at Yale

University, where he helped plan and lead part of the meeting devoted to evaluation of the PI's research results.

October 1993: EKSL organized and prsented the "CLIP/CLASP Workshop" hosted by Rome Laboratory, Griffiss AFB, for Planning Initiative participants. The workshop included:

- Two tutorials conducted by Paul Cohen (see "Tutorials," below).
- A presentation by David Westbrook on the use of CLIP.
- A technical introduction to CLASP presented by Adam Carlson.

July 1993: Paul Cohen organized and moderated a panel entitled "The Pros and Cons of Evaluation" at the *Eleventh National Conference on Artificial Intelligence* in Washington, DC. The panel included Lynette Hirschman (Mitre Corp.), Drew McDermott (Yale Univ.), Bruce Porter (UTexas/Austin) and Charles Weems (UMass/Amherst).

5.10 Tutorials

October 1993: Paul Cohen presented two tutorials at the "CLIP/CLASP Workshop" held at Rome Laboratory, Griffiss AFB: "Statistics Short Course: A Tutorial on Exploratory Data Analysis and Hypothesis Testing," and "Tutorial in the Use of CLIP/CLASP."

July 1993: Paul Cohen and Bruce Porter (Univ. of Texas, Austin) presented a tutorial entitled "Experimental Methods for Evaluating AI Systems," at the *Eleventh National Conference on Artificial Intelligence,* Washington, DC.

6 Awards, Promotions, and Honors

Paul Cohen was promoted from Associate to Professor of Computer Science, September, 1995.

Robert St. Amant was appointed Assistant Professor of Computer Science at North Carolina State University, August 1996.

David Jensen, previously with the Office of Technology Assessment, joined the EKSL in 1995 as Research Scientist. His research interests are in the areas of knowledge discovery in databases, statistical model building, machine learning, and advanced data analysis.

Scott Anderson was appointed Assistant Professor of Computer Science at Spelman College, August, 1995.

Adele Howe was appointed Assistant Professor of Computer Science at Colorado State University, Fort Collins, CO in September, 1993.

Dawn Gregory was selected in 1995 as a recipient of a three-year National Science Foundation Graduate Fellowhip merit award to support her Doctoral work in knowledge-based system induction and automated experiment design.

J. Tim Oates was awarded a three-year DoD National Defense Science and Engineering Graduate Fellowship in 1995 to support his work and training in knowledge-based planning and scheduling, statistical learning algorithms and distributed, multi-agent retrieval. p Paul Cohen was elected Fellow of the American Association for Artificial Intelligence in 1993, and served as Councillor of that organization from 1991 to 1994.

7 Related Service to the ARPI and the AI Research Community

Editorships

Paul Cohen and Bruce Porter edited a Special Issue of *Artificial Intelligence* devoted to empirical AI, published in August 1996.

Adele Howe currently serves as Book Review Editor for *AI Magazine*, and Associate Editor for North America for *Knowledge Engineering Review.*

Textbook on Empirical Methods. Paul Cohen's textbook, *Empirical Methods for Artificial Intelligence*, was published by the MIT Press in the summer of 1995. The book includes case studies from our work with TRANSSIM and PHOENIX, as well as other current and "classic" AI systems. It is distributed with CLIP/CLASP, an Instructor Guide, and a package of sample datasets that can be easily incorporated into a graduate level course on empirical methods. This text is an outgrowth of the ARPA/NSF-sponsored

Workshop on AI Methodology [1] that we hosted in 1990 in Northampton, MA.

AAAI-95 Spring Symposium on Integrated Planning Applications. Adele Howe (subcontractor at Colorado State Univ.) chaired a Spring Symposium on Integrated Planning Applications, whose purpose was to investigate issues that arise when a planning system is integrated into a real-world environment. Howe also chaired a panel on evaluation during that symposium.

Handbook of Evaluation for the ARPI. Paul Cohen helped compile the *Handbook of Evaluation for the ARPA/RL Planning Initiative.* The *Handbook* is based on discussions held at the December 93 Quarterly ARPI Meeting at Yale and includes contributions from Tom Dean, Yolanda Gil, Matt Ginsberg, Lou Hoebel and Cohen. Cohen made a panel presentation of the *Handbook* at the ARPI Tucson Workshop in 1994.

AAAI Panel on Evaluation. Paul Cohen organized and moderated a AAAI-93 panel on evaluation methods for AI that included Drew McDermott (Yale), Lynette Hirschman (Mitre Corporation), Bruce Porter (University of Texas) and Charles Weems (University of Massachusetts).

Article on Simulation and Empirical Evaluation. Paul Cohen was co-author of a three-way discussion of the merits and pitfalls of empirical evaluation using simulation testbeds that appeared in *AI Magazine*, winter 1993. This paper discussed some of the issues we have faced and advances we have made using empirical methods in the PHOENIX and TRANSSIM testbeds.

Program Committees. Paul Cohen served as a member of the Program Committees for the *Fifth* and *Sixth International Workshops on Artificial Intelligence and Statistics*, 1995 and 1997; the *Fourteenth National Conference on Artificial Intelligence* (AAAI-97); the *Third International Conference on Artificial Intelligence Planning Systems* (AIPS-96); the *Third International Workshop on Agent Theories, Architectures, and Languages* (ATAL-96); and the *First International Symposium on Intelligent Data Analysis* (IDA-95). He also served as Co-Chair of the Program Committee and Chair of the Steering Committee for IDA-97.

References

[1] Proceedings of the Workshop on Artificial Intelligence Methodology. Northampton MA, June 1991.

[2] Robert St. Amant and Paul R. Cohen. Preliminary system design for an EDA assistant. In *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 502–511, 1995. Also available as University of Massachusetts Computer Science Department Technical Report 94-79.

[3] Paul R. Cohen. A survey of the Eighth National Conference on Artificial Intelligence:

Pulling together or pulling apart? *AI Magazine*, 12(1):16–41, Spring 1991. Also available as University of Massachusetts Computer Science Department Technical Report 91-68.

[4] Paul R. Cohen. *Empirical Methods for Artificial Intelligence.* The MIT Press, Cambridge, 1995.

[5] Paul R. Cohen, Dawn E. Gregory, Lisa A. Ballesteros, and Robert St. Amant. Two algorithms for inducing structural equation models from data. In D. Fisher and H. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 3–12. Springer-Verlag, Inc., New York, 1996.

[6] Adele E. Howe, Robert St. Amant, and Paul R. Cohen. Integrating statistical methods for characterizing causal influences on planner behavior over time. In *Proceedings of the Sixth IEEE International Conference on Tools with AI*, pages 56–62. IEEE Computer Society Press, 1994. Computer Science Technical Report CS-94-115, Colorado State University.

[7] Adele E. Howe and G. Somlo. Modeling discrete event sequences as state transition diagrams. In *Advances in Intelligent Data Analysis: Reasoning About Data. Proceedings of IDA-97*, pages 573–584. SpringerVerlag, 1997.

[8] David Jensen. *Induction with Randomization Testing: Decision-Oriented Analysis of Large Data Sets.* PhD thesis, Washington University, 1992.

[9] George H. John. Robust decision trees: Removing outliers from databases. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 1995.

[10] Tim Oates. Fault identification in computer networks: A review and a new approach. Technical Report 95-113, University of Massachusetts at Amherst, Computer Science Department, 1995.

[11] Tim Oates and Paul R. Cohen. Searching for structure in multiple streams of data. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 346–354. Morgan Kaufmann Publishers, Inc., 1996.

[12] Tim Oates and David Jensen. The effects of training set size on decision tree complexity. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997.

[13] Tim Oates, Matthew D. Schmill, and Paul R. Cohen. Parallel and distributed search for structure in multivariate time series. "Submitted to the IEEE Transactions on Knowledge and Data Engineering Special Issue on Database Mining". Technical Report 96-23, University of Massachusetts Computer Science Department, 1996.

[14] Tim Oates, Matthew D. Schmill, Dawn E. Gregory, and Paul R. Cohen. Detecting complex dependencies in categorical data. In D. Fisher and H. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics*, pages 185–195. Springer Verlag, Inc., New York, 1996.

[15] J. Ross Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27:221–234, 1987.

[16] Dawn Gregory Tim Oates and Paul R. Cohen. Detecting complex dependencies in categorical data. In *Preliminary Papers of the Fifth International Workshop on AI and Statistics*, pages 417–423, 1995. Also available as University of Massachusetts Computer Science Department Technical Report 94-81.

☆U.S. GOVERNMENT PRINTING OFFICE: 1998-610-130-61220

DISTRIBUTION LIST

addresses	number of copies
WAYNE A. BOSCO AFRL/IFTB 525 BROOKS ROAD ROME, NY 13441-4505	2
EXPERIMENTAL KNOWLEDGE SYS LAB COMPUTER SCIENCE DEPT BOX 34610 UNIV OF MASSACHUSETTS AMHERST, MA 01003-9610	3
AFRL/IFOIL TECHNICAL LIBRARY 26 ELECTRONIC PKY ROME NY 13441-4514	1
ATTENTION: DTIC-OCC DEFENSE TECHNICAL INFO CENTER 8725 JOHN J. KINGMAN ROAD, STE 0944 FT. BELVOIR, VA 22060-6218	2
ADVANCED RESEARCH PROJECTS AGENCY 3701 NORTH FAIRFAX DRIVE ARLINGTON VA 22203-1714	1
DR JAMES ALLEN COMPUTER SCIENCE DEPT/BLDG RM 732 UNIV OF ROCHESTER WILSON BLVD ROCHESTER NY 14627	1
DR YIGAL ARENS USC-ISI 4676 ADMIRALTY WAY MARINA DEL RAY CA 90292	1
DR MARIE A. BIENKOWSKI SRI INTERNATIONAL 333 RAVENSWOOD AVE/EK 337 MENLO PRK CA 94025	1

DR MARK S. BODDY 1
HONEYWELL SYSTEMS & RSCH CENTER
3660 TECHNOLOGY DRIVE
MINNEAPOLIS MN 55418

DR MARK BURSTEIN 1
BBN SYSTEMS & TECHNOLOGIES
10 MOULTON STREET
CAMBRIDGE MA 02138

DR THOMAS L. DEAN 1
BROWN UNIVERSITY
DEPT OF COMPUTER SCIENCE
P.O. BOX 1910
PROVIDENCE RI 02912

DR PAUL R. COHEN 1
UNIV OF MASSACHUSETTS
COINS DEPT
LEDERLE GRC
AMHERST MA 01003

DR JON DOYLE 1
LABORATORY FOR COMPUTER SCIENCE
MASS INSTITUTE OF TECHNOLOGY
545 TECHNOLOGY SQUARE
CAMBRIDGE MA 02139

DR MICHAEL FEHLING 1
STANFORD UNIVERSITY
ENGINEERING ECO SYSTEMS
STANFORD CA 94305

RICK HAYES-ROTH 1
CIMFLEX-TEKNOWLEDGE
1810 EMBARCADERO RD
PALO ALTO CA 94303

DR JIM HENDLER 1
UNIV OF MARYLAND
DEPT OF COMPUTER SCIENCE
COLLEGE PARK MD 20742

MR. MARK A. HOFFMAN 1
ISX CORPORATION
1165 NORTHCHASE PARKWAY
MARIETTA GA 30067

DR RON LARSEN 1
NAVAL CMD, CONTROL & OCEAN SUR CTR
RESEARCH, DEVELOP, TEST & EVAL DIV
CODE 444
SAN DIEGO CA 92152-5000

DR. ALAN MEYROWITZ 1
NAVAL RESEARCH LABORATORY/CODE 5510
4555 OVERLOOK AVE
WASH DC 20375

ALICE MULVEHILL 1
BBN
10 MOULTON STREET
CAMBRIDGE MA 02238

DR DREW MCDERMOTT 1
YALE COMPUTER SCIENCE DEPT
P.O. BOX 2158, YALE STATION
51 PROPSPECT STREET
MEW HAVEN CT 06520

DR DOUGLAS SMITH 1
KESTREL INSTITUTE
3260 HILLVIEW AVE
PALO ALTO CA 94304

DR. AUSTIN TATE 1
AI APPLICATIONS INSTITUTE
UNIV OF EDINBURGH
80 SOUTH BRIDGE
EDINBURGH EH1 1HN - SCOTLAND

DIRECTOR 1
DARPA/ITO
3701 N. FAIRFAX DR., 7TH FL
ARLINGTON VA 22209-1714

DR STEPHEN F. SMITH 1
ROBOTICS INSTITUTE/CMU
SCHENLEY PRK
PITTSBURGH PA 15213

DR JONATHAN P. STILLMAN 1
GENERAL ELECTRIC CRD
1 RIVER RD, RM K1-5C31A
P. O. BOX 8
SCHENECTADY NY 12345

DR EDWARD C.T. WALKER 1
BBN SYSTEMS & TECHNOLOGIES
10 MOULTON STREET
CAMBRIDGE MA 02138

DR BILL SWARTOUT 1
USC/ISI
4676 ADMIRALTY WAY
MARINA DEL RAY CA 90292

DR MATTHEW L. GINSBERG 1
CIRL, 1269
UNIVERSITY OF OREGON
EUGENE OR 97403

MR JEFF GROSSMAN, CO 1
NCCOSC RDTE DIV 44
5370 SILVERGATE AVE, ROOM 1405
SAN DIEGO CA 92152-5146

DR ADELE E. HOWE 1
COMPUTER SCIENCE DEPT
COLORADO STATE UNIVERSITY
FORT COLLINS CO 80523

DR LESLIE PACK KAELBLING 1
COMPUTER SCIENCE DEPT
BROWN UNIVERSITY
PROVIDENCE RI 02912

DR SUBBARAO KAMBHAMPATI 1
DEPT OF COMPUTER SCIENCE
ARIZONA STATE UNIVERSITY
TEMPE AZ 85287-5406

MR DONALD P. MCKAY 1
PARAMAX/UNISYS
P O BOX 517
PAOLI PA 19301

DR MARTHA E POLLACK 1
DEPT OF COMPUTER SCIENCE
UNIVERSITY OF PITTSBURGH
PITTSBURGH PA 15260

DR MANUELA VELOSO 1
CARNEGIE MELLON UNIVERSITY
SCHOOL OF COMPUTER SCIENCE
PITTSBURGH PA 15213-3891

DR DAN WELD 1
DEPT OF COMPUTER SCIENCE & ENG
MAIL STOP FR-35
UNIVERSITY OF WASHINGTON
SEATTLE WA 98195

DR TOM GARVEY 1
DARPA/ISO
3701 NORTH FAIRFAX DRIVE
ARLINGTON VA 22203-1714

DIRECTOR 1
ARPA/ISO
3701 NORTH FAIRFAX DRIVE
ARLINGTON VA 22203-1714

OFFICE OF THE CHIEF OF NAVAL RSCH 1
ATTN: MR PAUL QUINN
CODE 311
800 N. QUINCY STREET
ARLINGTON VA 22217

DR GEORGE FERGUSON 1
UNIVERSITY OF ROCHESTER
COMPUTER STUDIES BLDG, RM 732
WILSON BLVD
ROCHESTER NY 14627

DR STEVE HANKS 1
DEPT OF COMPUTER SCIENCE & ENG'G
UNIVERSITY OF WASHINGTON
SEATTLE WA 98195

DR ADNAN DARWICHE 1
INFORMATION & DECISION SCIENCES
ROCKWELL INT'L SCIENCE CENTER
1049 CAMINO DOS RIOS
THOUSAND OAKS CA 91360

MR.ROBERT J. KRUCHTEN 1
HQ AMC/SCA
203 W LOSEY ST, SUITE 1016
SCOTT AFB IL 62225-5223

DR. MAREK RUSINKIEWICZ 1
MICROELECTRONCS & COMPUTER TECH
3500 WEST BALCONES CENTER DRIVE
AUSTIN, TX 78759-6509

MAJOR DOUGLAS DYER/ISO 1
DEFENSE ADVANCED PROJECT AGENCY
3701 NORTH FAIRFAX DRIVE
ARLINGTON, VA 22203-1714

DR. STEVE LITTLE 1
MAYA DESIGN GROUP
2100 WHARTON STREET S&E 702
PITTSBURGH, PA 15203-1944

NEAL GLASSMAN 1
AFOSR
110 DUNCAN AVENUE
BOLLING AFB, WASHINGTON, D.C.
29332